SUBJECT IDENTIFICATION USING 3D FACIAL LANDMARKS

Sk Rahatul Jannat, Diego Fabiano, and Shaun Canavan

University of South Florida

ABSTRACT

Landmark localization is an important first step towards geometric based vision research including subject identification. Considering this, we propose to use 3D facial landmarks for the task of subject identification. To detect the 3D landmarks, we propose the use of a modified version of the Temporal Deformable Shape Model. We show that the detected 3D facial features can be used to model a wide range of subject identities, including those with large variations in expression. Experiments are conducted using a Support Vector Machine (SVM), Random Forest (RF), and Long Short-term Memory (LSTM) neural network for identification, on the BU-4DFE, BP4D, and BP4D+ 3D/4D face databases. We show that our proposed method outperforms current state of the art methods for subject identification on BU-4DFE and BP4D. To the best of our knowledge, this is the first work to investigate subject identification on the BP4D+, resulting in a baseline for the community.

Index Terms— subjection identification, 3D, SVM, random forest, long short-term memory neural network

1. INTRODUCTION

Broadly, face recognition can be categorized as holistic, hybrid matching, or feature-based [39]. Holistic approaches look at the global similarity of the face such as a 3D morphable model (3DMM) [2]; hybrid matching make use of either multiple methods [14] or multiple modalities [17]; feature-based methods look at local features of the face to find similarities [40]. The work proposed in this paper can be categorized as feature-based. Due to its non-intrusive nature and wide applicability in security and defense related fields, face recognition has been actively researched by many groups in recent decades.

Since some of the earlier methods for face recognition [30], [38], to more recent works within the past 10 years [5], [34] 2D face recognition has been an actively researched field. With the recent advances in deep neural networks, we have seen significant jumps in performance [11], [18], [22], [24], [26], [32]. Liu et al. [21] proposed the angular softmax that allows convolutional neural networks (CNN) the ability to learn angularly discriminative features. This was proposed to handle the problem where face features are shown to have a smaller intra-class distance compared to inter-class distance. Recently, Tran et al. [29] proposed regressing 3D morphable model shape and texture parameters from a 2D image using a CNN. Using this approach, they were able to obtain a sufficient amount of training data for their network showing promising results. Zhu et al. [41] proposed a high-fidelity pose and expression normalization method that made use of a 3DMM to generate natural, frontal facing, neutral face images. Using this method, they achieved promising results in both constrained and unconstrained environments (i.e. wild settings). Although performance has been increasing and groups have been actively working on 2D issues such as pose and lighting, there are still some challenges that occur. In recent years, there has been more of an interest in using 3D face recognition to solve these issues [10], [11], [25] due to the development of powerful, high-fidelity 3D sensors.

Echeagaray-Patron et al. [10] proposed a method for 3D face recognition where conformal mapping is used to map the original face surfaces onto a Riemannian manifold. From the conformal and isometric invariants that they compute, comparisons are then made. This method was shown to be have invariance to both expression and pose. Li et al. [20] proposed the use of SIFT-like matching using three 3D key point descriptors. Each of these descriptors were fused at the feature-level to describe local shapes of detected key points. Lei et al. [19] proposed the Angular Radial Signature for 3D face recognition. This signature is extracted from the semi-rigid regions of the face, followed by mid-level features being extracted from the signature by Kernel Principal Component Analysis. These features were then used to train a support vector machine showing promising results when comparing neutral vs. non-neutral faces. Berretti et al. [1] proposed the use of 3D Weighted Walkthroughs with iso-geodesic facial strips for the task of 3D face recognition. They achieved promising results on the FRGC v2.0 [23] and SHREC08 [8] 3D facial datasets. Using multistage hybrid alignment algorithms and an annotated face model, Kakadiaris et al. [15] used a deformable model framework to show robustness to facial expressions when performing 3D face recognition.

Motivated by these works, we propose the use of a modified version of the Temporal Deformable Shape Model [6], to detect 3D facial landmarks for subject identification. See Figure 1 for an overview of the proposed approach. The main contribution of this work is 3-fold and is summarized as follows:

1. Propose the use of 3D facial landmarks for subject

identification using an SVM, RF, and LSTM network.
2. Test the efficacy of our method on 3 publicly available 3D faces databases [33], [35], [36], [37]. Combined, we test our method on over 620,000 3D faces across these 3 databases and show state of the art results on BU-4DFE [33] and BP4D [35], [36].
3. To the best of our knowledge this is the first work to perform subject identification on the BP4D+ [37] 3D face database, detailing a baseline for the community.

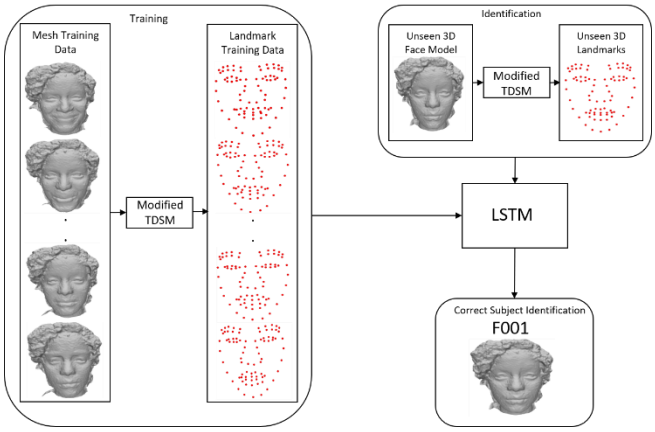

Figure 1. Overview of proposed method for subject identification. Example is showing an unseen 3D mesh model of subject 'F001' from BP4D+ [37], who is correctly identified based on training a LSTM [13] from 3D facial data detected from the proposed SDDM.

2. MODIFIED TEMPORAL DEFORMABLE SHAPE MODEL

2.1. Temporal deformable shape model

The Temporal Deformable Shape Model (TDSM) [6] models the shape variation of 3D facial data. Given a sequence of data (i.e. 4D), it also models the implicit constraints on shape that are imposed (e.g. small changes in motion and shape). To construct a TDSM a training set of 3D facial landmarks is required. First, the 3D facial landmarks are aligned using a modified version of Procrustes analysis [9]. Given a training set of size *L* 3D faces, where each face has *N* facial landmarks (aligned with Procrustes analysis), a parameterized model *S* is constructed, where $S = F_1^1,\ldots, F_N^1, \ldots, F_1^m,\ldots, F_N^m$. F_i^m is the i^{th} landmark of the m^{th} 3D face in the training set, where $F_i^m = (x_i^m, y_i^m, z_i^m)$ and $1 \leq m \leq L$. From this model, principal component analysis (PCA), is then applied to learn the modes of variation, *V*, of the training data.

Given the parameterized model, *S*, and the modes of variation, *V*, to detect 3D facial landmarks, an offline weight vector, *w*, is constructed that allows for new face shapes to be constructed (i.e. these face shapes are constructed offline), by a linear combination of landmarks as $S = \bar{s} + Vw$, where $\bar{s}$ is the average face shape. These constructed face shapes are constrained to be within the range $-2\sqrt{\lambda_i} \leq w_i \leq 2\sqrt{\lambda_i}$, where w_i is the i^{th} weight in the range, and λ_i is the i^{th} eigenvalue from PCA. This constraint is imposed to make sure the new face shape is a 3D face.

To fit (i.e. detect landmarks) to a new input mesh, the offline table of weights (*w*) is constructed with a uniform amount of variance. The Procrustes distance, *D*, is then computed between each face shape (referred to as an instance of the TDSM) and the new input mesh. The smallest distance is considered the best detected landmarks. This is not meant to be an exhaustive overview of a TDSM, therefore we refer the reader to the original work [6] for more details.

2.2. Modified temporal deformable shape model

The TDSM has no direct convergence criterion as it calculates *D* for all instances that have been created. Considering this, we propose a modified version with a convergence criterion. For this version, model construction is carried out in the same manner, however, the detection of 3D landmarks is optimized. We modify this part of the algorithm by proposing a convergence criterion based on the mean squared error (MSE) calculated between the new detected points (found from the Procrustes distance), and the original instance that was used. Given detected landmarks on an input mesh, the MSE is then calculated between these landmarks, and the landmarks from the instance that was used to detect them on the input mesh. This can be done, as the MSE will be low if a good fit has been found, otherwise it will be high.

To find the best detected landmarks, *w* is varied by stepping through the range $[-\sigma, \sigma]$ (σ is the standard deviation of the training set) and constructing new instances to detect landmarks on the input mesh model. Starting at $w = 0$ (i.e. the average face), we make one step in a positive direction towards σ , and one in a negative direction towards -σ. Both instances are then used to detect landmarks on the input mesh, and their MSE is calculated. Whichever instance results in the lowest MSE is chosen as the starting point for the first iteration. We then continue to step in the same direction of the starting instance (either towards σ or $-\sigma$) until the MSE increases. Once this occurs, we keep the previous iteration as we have found the global minimum of the TDSM detection process resulting in the best landmarks. See Figure 2 for examples of the detection process, including the MSE for each step.

3. EXPERIMENTAL DESIGN AND RESULTS

Using the modified TDSM algorithm, detailed in section 2, we detected 83 facial landmarks on 3 publicly available 3D face databases: BU4DFE [33], BP4D [35][36], and BP4D+ [37]. From these facial landmarks, we then conducted subject identification experiments, where the landmarks are used as training data for 3 machine learning classifiers (detailed in 3.2). Using these 83 facial landmarks we have also reduced the dimensionality of the 3D faces from over 30,000 3D vertices, while still retaining important features for subject identification. This allows us to reduce storage requirements,

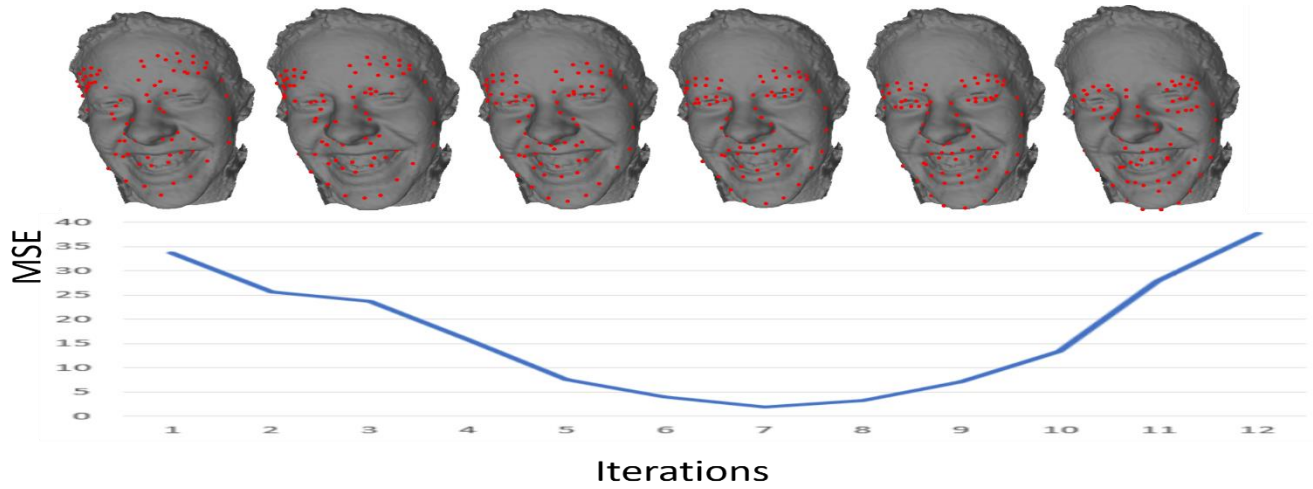

Figure 2. 12 steps (iterations) of a TDSM on a 3D model from BP4D+. Top row shows detected landmarks for steps 1, 3, 5, 7, 9, and 11 (from left to right). Bottom row shows graph of MSE for each step. Iteration 7 has lowest MSE of 1.89 and results in the best detected landmarks. *NOTE: More iterations shown for visual purposes only as the modified TDSM converged at iteration 7. Best viewed in color.*

as well as processing time of the 3D face, which can be limitations of 3D face recognition [4], [16]. An overview of the databases and the experimental design is detailed in the following subsections.

3.1. 3D face databases

For our experiments, we chose 3 state-of-the-art 3D face databases, and investigate a total of 620,326 3D facial landmarks (i.e. faces). Details on each are given below.

BU-4DFE [33]: Consists of 101 subjects displaying 6 prototypic facial expressions plus neutral. The dataset consists of 58 females and 43 males, including a variety of racial ancestries. The age range of the BU-4DFE is 18-45 years of age. In total, there are over 60,000 frame of 3D facial data; which we used all of for our experiments.

BP4D [35], [36]: Consists of 41 subjects displaying 8 expressions plus neutral. It consists of 23 females and 18 males; 11 Asian, 4 Hispanic, 6 African-American, and 20 Euro-American ethnicities are represented. The age range of the BP4D is 18-29 years of age. For our experiments, we again used the entire database, which consists of over 360,00 frames of 3D facial data. Although this database was developed to explore spatiotemporal features in facial expressions, due to its size and large variation in expression it is a natural fit for our subject identification study.

BP4D+ [37]: Consists of 140 subjects (82 females and 58 males) ages 18-66. This data corpus consists of ethnic and racial ancestries that include Black, White, and Asian each with highly varied emotions. These emotions are elicited through tasks designed to elicit dynamic emotions in the subjects such as disgust, embarrassment, pain, and surprise resulting in a challenging dataset. For our experiments, we selected a subset of the entire database (over 1.5 million frames), which consists of over 150,000 frames of data distributed across all 140 subjects. Like the BP4D database, this was designed to study emotion classification, however, due to the diversity of subjects, large variety and range of emotions, it too is a natural fit for our study.

3.2. Experimental design

To conduct our experiments, we detected 83 facial landmarks on the 3D data using the modified TDSM (section 2.2.). Given 3D facial landmarks, we then translate them so that the centroid of the face is located at the origin in 3D space, to align the data. The translated 3D facial features are then used for subject identification. Each of the 3D facial landmarks (x, y, z coordinates) are inserted into a new feature vector. For all 83 landmarks this gives us a feature vector of size $83 \times 3 = 249$. This feature vector is used to train the classifiers for subject identification. For our experiments we trained a support vector machine (SVM) [31], random forest (RF) [3], and Long short-term memory (LSTM) neural network [13]. Our network consists of 1 short-term memory layer with a look back of 2 faces (estimated landmarks), followed by 0.5 dropout, and a fully connected layer for classification. The softmax activation function was used, along with the RMSprop [28] optimizer with a learning rate of 0.0001.

For each classifier, each subject's identity was used as the class (each 3D face is labeled with a subject id). As we will show; accurate results on an SVM, RF, and LSTM show the robustness of the 3D facial landmarks to multiple machine learning classifiers. Using the 3D face databases detailed in 3.1, we used 10-fold cross validation for training and testing for our subject identification experiments. The data is randomly split into 10 subsets where one set is used for testing and the other nine are used for training. Each of the

subsets is used for testing, where each iteration separates the test set from the training data.

3.3. Subject identification results

Using the experimental design detailed in the previous section, we achieved an average correct subject identification rate of 99.91%, 99.96%, and 99.93% for a RF, SVM, and LSTM respectively, across all databases (Table 1).

Table 1. Subject identification accuracies for the 3 tested datasets and classifiers.

	BU4DFE	BP4D	BP4D+
SVM	99.9%	**99.9%**	**99.9%**
RF	**100%**	**99.9%**	98.8%
LSTM	**100%**	**99.9%**	**99.9%**

As can be seen in table 1, an SVM, RF, and LSTM can accurately identify subjects from the BU4DFE, BP4D, and BP4D+ datasets achieving a max accuracy of 100% on the BU4DFE, and a minimum accuracy of 98.8 percent on the BP4D+. All three of the tested classifiers achieved consistent results across all three datasets. As each of the datasets, contains large variations in expression, these results suggest the detected 3D landmarks are invariant to expression changes, for the task of subject identification.

3.4. Comparisons to state of the art

We also compared our proposed method to the current state of the art on BU-4DFE [33] and BP4D [35], [36]. As previously mentioned, to the best of our knowledge this is the first study to perform subject identification on BP4D+ [37]; therefore, we did not have any works to compare against resulting in a baseline for the community. Also, to the best of our knowledge, there is only one prior work detailing subject identification results on the BP4D dataset as shown below. The comparisons for BU-4DFE and BP4D can be seen in Table 2.

Table 2. Comparisons of proposed method to current state of the art on BU-4DFE and BP4D.

Method	BU-4DFE	BP4D
Proposed method (RF)	**100%**	**99.9%**
Proposed method (LSTM)	**100%**	**99.9%**
Proposed method (SVM)	99.9%	**99.9%**
Sun et al. [27]	98.61%	NA
Fernandes et al. [12]	96.71%	NA
Canavan et al. [7]	92.7%	93.4%

It is important to note that Canavan et al [7] used a small subset of both the BU-4DFE and BP4D datasets for their experiments. They used 1800 and 2400 respectively, while we used all data in both datasets (60402 and 367474 respectively). The work from Sun et al. [27] also requires both spatial and temporal information to achieve their results of 98.61%, while our approach can identify a subject based on one frame of data, which is useful when temporal information does not exist.

4. CONCLUSIONS

We have proposed detecting 3D facial landmarks with a modified TDSM for the task of subject identification. We have shown that detected landmarks can be used to accurately identify 282 subjects from three 3D face databases, for a total of 620,326 faces. The proposed method outperforms current state of the art on 2 publicly available 3D face databases achieving a max identification accuracy of 100% on BU-4DFE and 99.9% on BP4D. To the best of our knowledge, this is the first work to report subject identification results on the BP4D+, resulting in a new baseline for the community.

Using the detected facial landmarks, we have detailed accurate subject identification results using a random forest, support vector machine, and long short-term neural networks. We have also shown that the detected 3D facial landmarks can decrease the overall storage requirements of 3D facial data, while maintaining robustness to multiple machine learning classifiers, as well as robustness to facial expression. We will investigate this robustness/invariance to expression in future work, by investigating the entire BP4D+ (compared to the subset detailed here), as well as other state-of-the-art 3D face datasets that contain expressions.

5. REFERENCES

[1] S. Berretti A. Del Bimbo, and P. Pala, "3D Face Recognition Using Isogeodesic Strips," IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(12): 2162-2177, 2010.

[2] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(9): 1063-1074, 2003.

[3] L. Breiman, "Random forests," *Machine Learning*, 45(1):5-32, 2001.

[4] K. Bowyer, K. Chang, and P. Flynn, "A Survey of Approached and Challenges in 3D and Multimodal 3D+2D Face Recognition," Computer Vision and Image Understanding, 101(1): 1-15, 2006.

[5] S. Canavan, B. Johnson, M. Reale, Y. Zhang, L. Yin, and J. Sullins, "Evaluation of multi-frame fusion based face classification under shadow," International Conference on Pattern Recognitino, pp. 1265-1268, 2010.

[6] S. Canavan, X. Zhang, and L. Yin, "Fitting and tracking 3D/4D facial data using a temporal deformable shape model," International Conference on Multimedia and Expo, 2013.

[7] S. Canavan, P. Liu, X. Zhang, and L. Yin, "Landmark localization on 3D/4D range data using a shape index-based statistical shape model with global and local constraints," Computer Vision and Image Understanding, 2015.

[8] M. Daoudi, F. Haar, and R. Veltkamp, "SHREC 2008 – Shape Retrievel Contest of 3D Face Scans," http://give-lab.cs.uu.nl/SHREC/ shrec2008/, 2008.

[9] M. de Bruijne, B. Van Ginneken, M. Viergever, and W. Niessen, "Adapting shape models for 3D semgentation of tubular structure in medical images," International Conference on Information Processing in Medical Imaging, pp. 136-147, 2013.

[10] B. A. Echeagarray-Patron, V.I. Kober, V.N. Karnaukhov, and V. Kuznetsov, “A method of face recognition using 3D facial surfaces,” Journal of Communications Technology and Electronics, 62(6): 648-652, 2017.

[11] M. Emambakhsh, and A Evans, “Nasal patches and curves for expression-robust 3D face recognition,” IEEE PAMI, 39(5): 995-1007, 2017.

[12] S. Fernandes, and G. Bala. "3D and 4D face recognition: a comprehensive review." Recent Patents on Engineering, 8(2): 112-119, 2014.

[13] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, 1997.

[14] J. Huang, B. Heisele, and V. Blanz, “Compont-based face recognition with 3D morphable models,” International Conference on Audio and Video-based Person Authentication, 2003.

[15] I. Kakadiaris, G. Passalis, G. Toderici, N. Murtuza, and T. Theoharis, “3D Face Recognition,” British Machine Vision Conference, 2006.

[16] I. Kakadiaris, G. Passalis, G. Toderick, M. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis, “Three-dimensional Face Recognition in the Presence of Facial Expressions: An Annoated Deformable Model Approach,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4): 640-649, 2007.

[17] I. Kakadiaris, G. Passalis, T. Theoharis, G. Konstantinidis, and N. Murtuza, “Multimodal face recognition: combination of geometry with physiological information,” Computer Vision and Pattern Recognition, 2005.

[18] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, “The megaface benchmark: 1 million faces for recognition at scale,” CVPR, 2016.

[19] Y. Lei, M. Bennamoun, M. Hayat, and Y. Guo, “An efficient 3D face recognition approach using local geometrical signatures,” Pattern Recognition, 47(2): 509-524, 2014.

[20] H. Li, D. Huang, J. Morvan, Y. Wang, and L. Chen, “Towards 3d face recognition in the real: a registratino-free approach using fine-grained matching of 3D kepoint descriptors,” International Journal of Computer Vision, 113(2): 128-142, 2015.

[21] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: deep hypershpere embedding for face recognition,” Computer Vision and Pattern Recognition, 2017.

[22] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” British Machine Vision Conference, 2015.

[23] P. Philips, P. lynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, “Overview of the Face Recognition Grand Challenge,” Computer Vision and Pattern Recognition, pp. 947-954, 2005.

[24] M. Saragih, S. Lucey, and J. Cohn, “Deformable model fitting by regularized landmark-mean-shift,” International Journal of Computer Vision, 91(2): 200-215, 2011.

[25] S. Sima, B. Boubakeur, and Q.M.J. Wu, “A survey of local feature methods for 3D face recognition,” Pattern Recognition, 72, pp. 391-406, 2017.

[26] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” NIPS, 2014.

[27] Y. Sun and L. Yin, "3D Spatio-Temporal face recognition using dynamic range model sequences." Computer Vision and Pattern Recognition Workshops, 2008.

[28] T. Tielman, and G. Hinton, “Rmsprop: divide the gradient by a running average and its recent magnitude,” COURSERA: Neural Networks for Machine Learning, 31, Technical Report, 2012.

[29] A. Tran, T. Hassner, I. Masi, and G. Medioni, “Regressing robust and discriminitive 3D morphable models with a very deep neural network,” Computer Vision and Pattern Recognition, 2017.

[30] M. Turk and A. Pentland, “Face recognition using eigenfaces,” Computer Vision and Pattern Recognition, 1991.

[31] V. Vapnik, “The support vector method of function estimation,” Nonlinear Modeling, pp.55-85, 1998.

[32] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” ECCV 2016.

[33] L. Yin, X. Chen, et al., “A high-resolution 3d dynamic facial expression database,” *Face and Gesture*, 2008.

[34] L. Zhang, M. Yang, X. Feng, “Sparse representation or collaborative representation: Which helps face recognition?” International Conference on Computer Vision, 2011.

[35] X. Zhang, L. Yin, et al., “BP4D-Spontaneous: A high resolution 3D dynamic facial expression database,” *Image and Vision Computing*, 32(10):692-706, 2014.

[36] X. Zhang, L. Yin, et al., “A high resolution spontaneous 3D dynamic facial expression database,” *Face and Gesture*, 2013.

[37] Z. Zhang, J. Girard, Y. Wu, et al., “Multimodal spontaneous emotion corpus for human behavior analysis,” Computer Vision and Pattern Recognition, 2016.

[38] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng, “Discriminant analysis of principal components for face recognition,” Face Recognition, pp.73-85, 1998.

[39] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, “Face recognition: a literature survey,” ACM Computing Surveys, 35(4): 399-458, 2003.

[40] C. Zhong, Z. Sun, and T. Tan, “Robust 3d face recognition using learned visual codebook,” Computer Vision and Pattern Recognition, 2007.

[41] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Li, “High-fidelity pose and expression normalization for face recognition in the wild,” Computer Vision and Pattern Recognition, 2015.